03 | 数据中台建设三板斧：方法论、组织和技术

2020-04-06 郭忆

数据中台实战课 进入课程 >

讲述：郭忆
时长 20:57　大小 19.20M

你好，我是郭忆。

在上一讲中，我带你了解了什么样的企业适合建数据中台，可能有的同学会说：你可真的戳中我了，我们现在就面临这个问题，可是知道要转型，要建设数据中台，却不知道要咋做，怎么办呢？

现在有很多讲“如何建设数据中台”的文章，大家的观点各不相同：

> 有的观点说，数据中台是一种数据建设的方法论，按照数据中台设计方法和规范实施就可以建成数据中台了；

也有观点认为，数据中台的背后是数据部门组织架构的变更，把原先分散的组织架构形成一个统一的中台部门，就建成了数据中台；

除此之外，你可能还听到过一些大数据公司说，他们可以卖支撑数据中台建设的产品技术。

那数据中台到底如何建设呢？咱们先不着急回答这个问题，而是看一个例子。

你肯定见过盖房子，盖房子之前，是不是先得有设计图纸，知道如何去盖这个房子？然后还必须要有一个好用的工具（比如水泥搅拌机、钢筋切割机）帮你盖好这个房子。当然了，盖房子离不开一个靠谱的施工队伍，这里面涉及很多角色（泥瓦工、木工、水电工等等），这些人必须高效协作，最终才能盖出一个好的房子。

如果我们把建数据中台比作是盖房子，那么设计图纸就是数据中台建设的方法论；工具是数据中台的支撑技术；施工队伍就是数据中台的组织架构。这三者缺一不可。

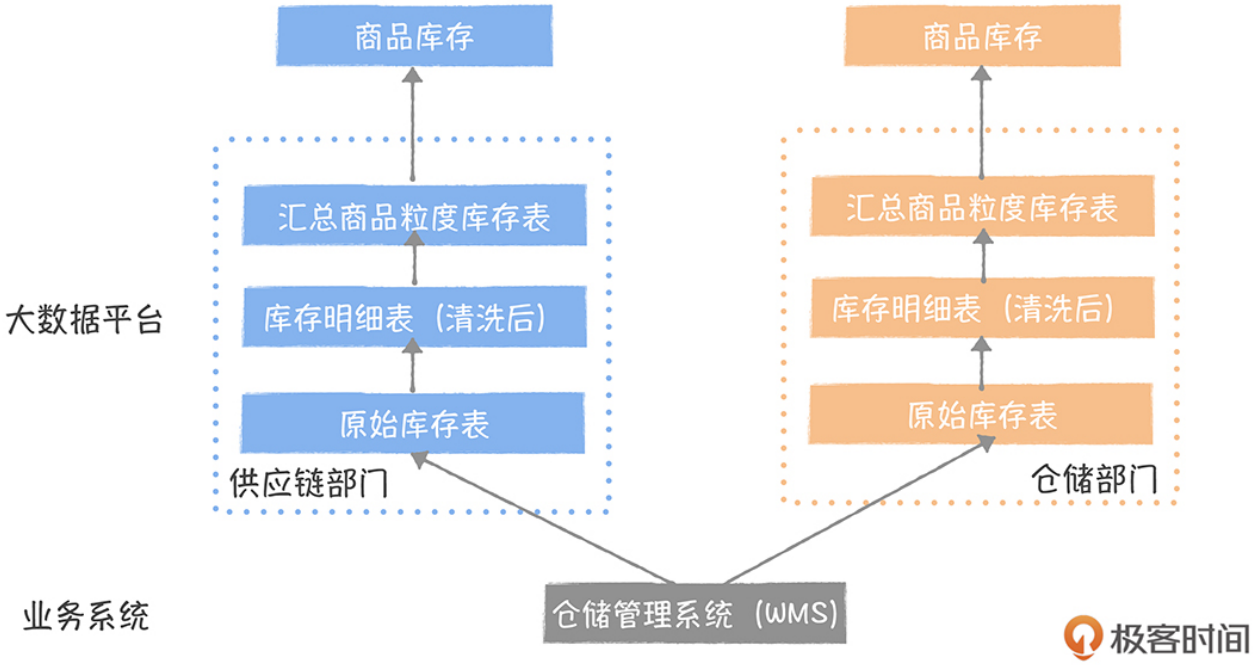
这一讲我就以全局的视角，让你从宏观上了解如何建设一个企业级的数据中台。

数据中台建设方法论

早在 2016 年，阿里巴巴就提出了数据中台建设的核心方法论：OneData 和 OneService。经过这么多年，很多公司都进行了实践，但你很难找出一个准确的定义去描述这些方法论，而我结合自己在网易数据中台建设的经验，对方法论进行了系统化的定义，你可以借鉴参考一下。

OneData

用一句话定义 OneData 的话，就是所有数据只加工一次。 这个概念具体是啥意思呢？我们来看一个例子。



电商业务建设数据中台之前，每个部门内部都会有一些小的数仓去完成本部门的数据分析需求。

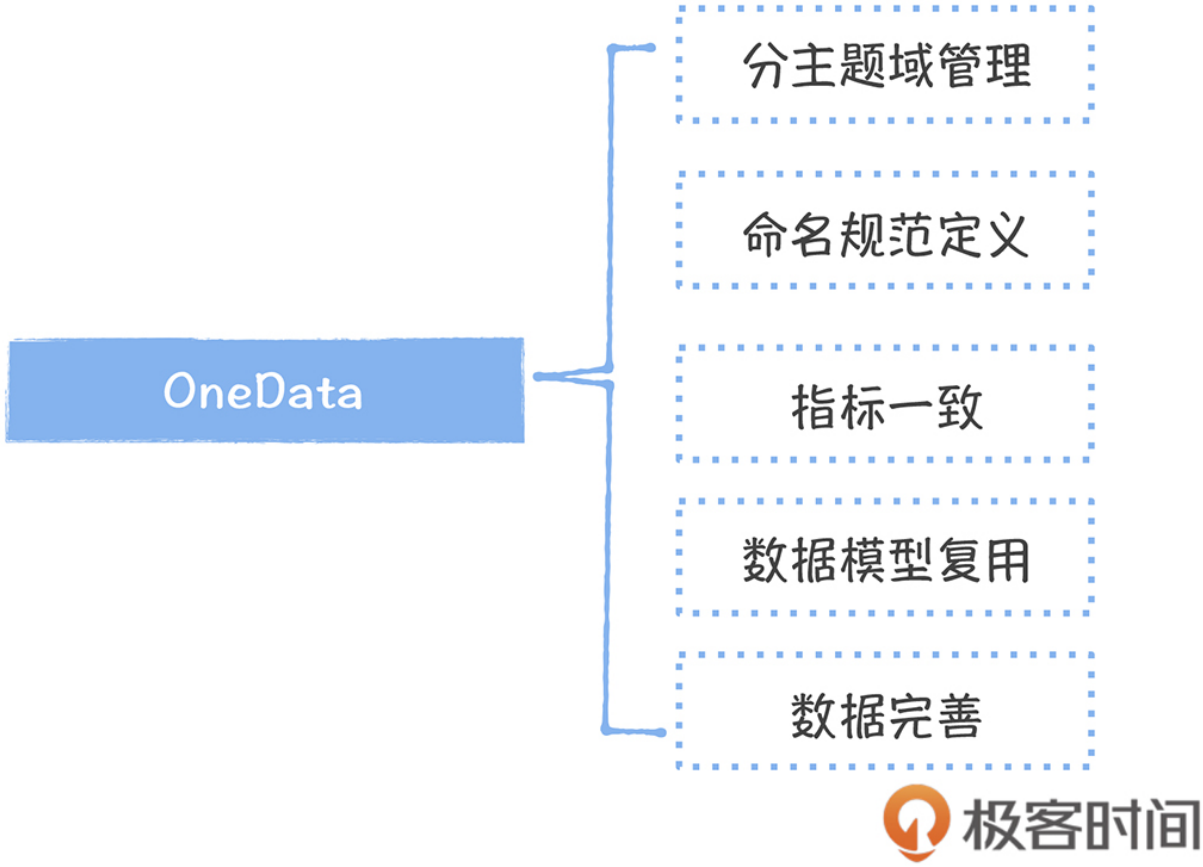
有一天，供应链团队接到一个数据需求，那就是计算“商品库存”指标，供应链的运营需要根据每个商品的库存制订商品采购计划，部门的数据开发从业务系统同步数据，进行数据清洗、聚合、深度加工，最终，产出这个指标花了 1 周的时间。

而恰逢全年最重要的大促节日，市场部门也需要根据每个商品的库存，制订商品的促销计划。该数据开发接到这个紧急的需求（与供应链团队类似）从需求开发到上线，也足足花费了 1 周的时间。同部门的运营会抱怨说，为什么数据需求开发这么慢，根本无法满足大促期间高频的市场运营决策。而对公司而言，等待 1 周意味着遭受巨大损失，该促销的商品没有促销，不该促销的却低价卖了。

如果你是这个公司的老板， 肯定会问，既然供应链团队已经计算出来了商品库存的数据，为什么市场部门不直接用，还要从头再计算一遍呢？这个看似很傻的行为，却处处出现在我们日常的数据建设中。

而数据中台就是要在整个电商业务形成一个公共数据层，消灭这些跨部门的小数仓，实现数据的复用，所以强调数据只加工一次，不会因为不同的应用场景，不同的部门数据重复加工。

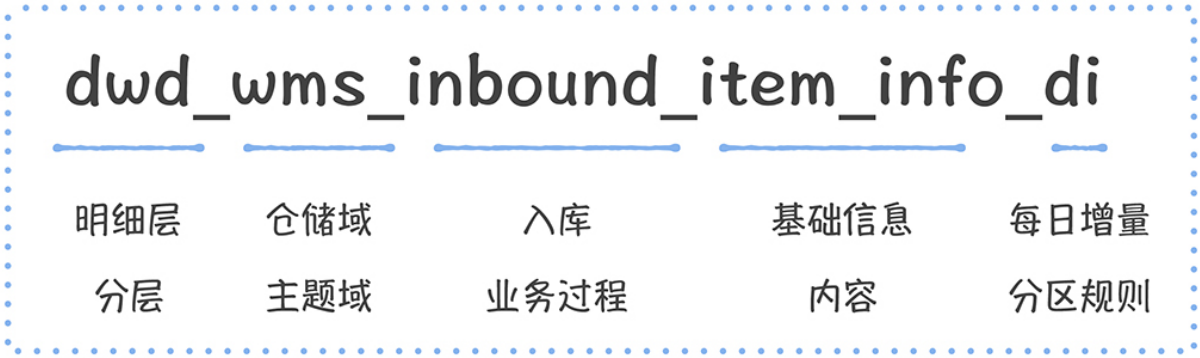
那具体来说，如何去做才能实现数据只加工一次呢？有这样五点：



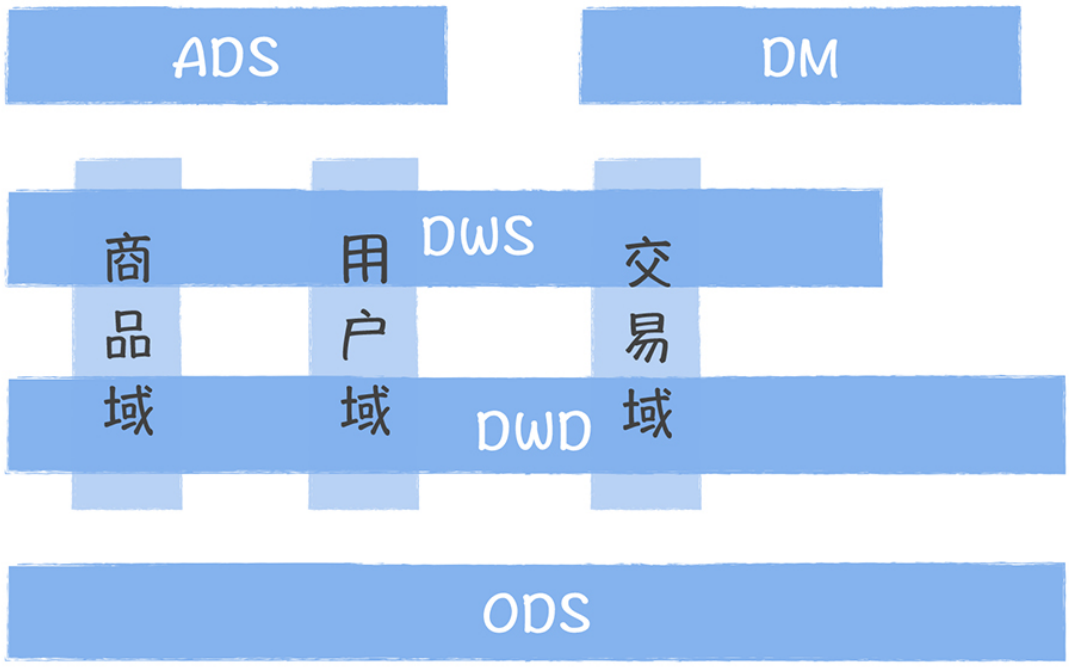
试想一下，现在你构建了数据中台，但存在几万张表，同时又有几十个数据开发维护这些表，如何来确保这些表的管理效率？ **我建议你选择划主题域。**

我们可以将这几万张表划到不同的主题域中，比如在电商业务中，商品、交易、流量、用户、售后、配送、供应链都可以作为主题域。好的主题域划分，是相对稳定，尽可能地覆盖绝大多数的表。

除此之外，还要对表的命名进行规范化统一， 表的名称中最好能够携带表的主题域、业务过程、分层以及分区信息。比如，对于仓储域下的一张入库明细表，规则命名可以这样：



接着你还必须构建全局的指标字典，确保所有表中相同指标的口径必须一致（这部分内容我会在 06 讲细说）。



另外，为了实现模型的复用，数据中台适合采用分层的设计方式，常见的分层包括：ODS 原始数据层，DWD 明细数据层，DWS 轻度汇总数据层，ADS/DM 应用数据层 / 数据集市层。**最后，数据中台的数据必须尽可能的覆盖所有的业务过程，**数据中台中每一层的数据也要尽可能完善，让数据使用者尽可能的使用汇总后的数据。

OneData 体系的目标是构建统一的数据规范标准，让数据成为一种资产，而不是成本。资产和成本的差别在于资产是可以沉淀的，是可以被复用的。成本是消耗性质的、是临时的、无法被复用的。

OneService

OneService，数据即服务，强调数据中台中的数据应该是通过 API 接口的方式被访问。

这里我想提个问题：为什么数据一定要通过 API 接口的方式被访问，不通过 API 接口，直接提供数据表给用户又存在哪些问题呢？

如果你是数据应用开发，当你要开发一个数据产品时，首先要把数据导出到不同的查询引擎上：

数据量小的使用 MySQL；

大的可能用到 HBase；

需要多维分析的可能需要 Greenplum；

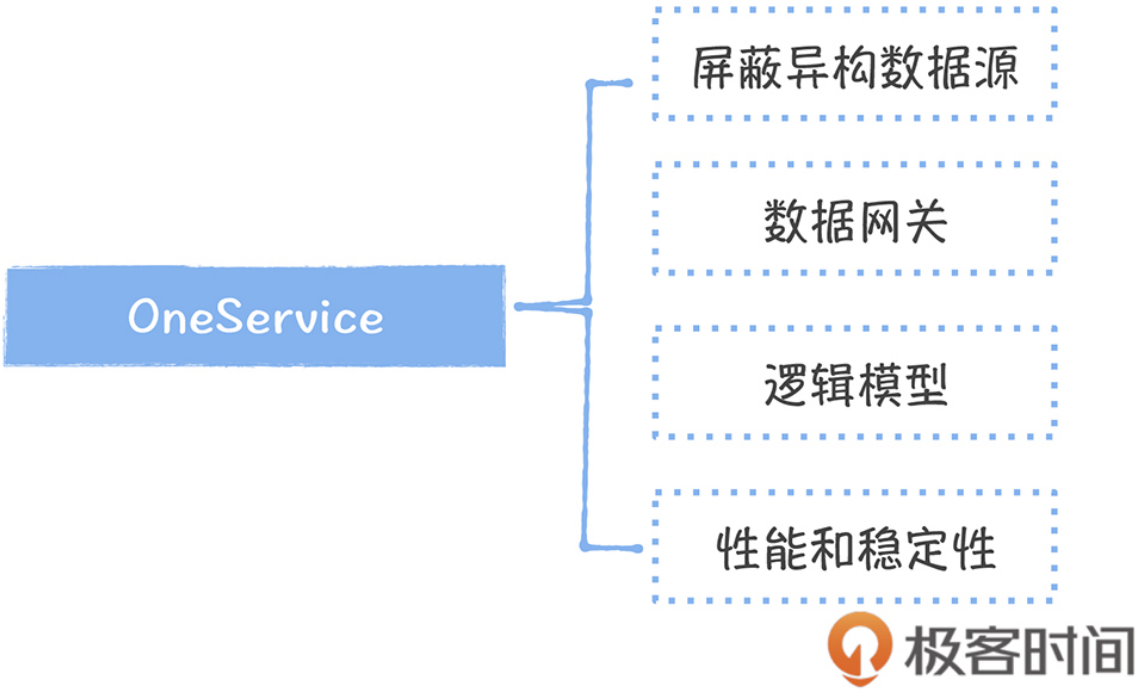
实时性要求高的需要用到 Redis；

总的来说，不同的查询引擎，应用开发需要定制不同的访问接口。

如果你是一个数据开发，当某个任务无法按时产出，发生异常时，想要了解这个表可能会影响到下游的哪些应用或者报表，但是却发现单纯依赖表与表的血缘无法触及应用，根本无法知道最后的这些表被哪些应用访问。与此同时，当你想下线一张表时，因为不知道谁访问了这张表，无法实施，最终造成了“上线容易，下线难”的窘境。

而 API 接口一方面对应用开发屏蔽了底层数据存储，使用统一标准的 API 接口查询数据，提高了数据接入的速度。另一方面，对于数据开发，提高了数据应用的管理效率，建立了表到应用的链路关系。

那如何来实现数据服务化呢？



屏蔽异构数据源：数据服务必须要能够支撑类型丰富的查询引擎，满足不同场景下数据的查询需求，常见的有 MySQL、HBase、Greenplum、Redis、Elasticsearch 等。

数据网关：要实现包括权限、监控、流控、日志在内的一系列管控能力，哪个应用的哪个页面访问了哪个模型，要做到实时跟踪，如果有一些模型长时间没有被访问，应该予以下线。使用数据的每个应用都应该通过 accesskey 和 secretkey 实现身份认证和接口权限的管理。另外，访问日志可以方便在访问出现问题时，加快排查速度。

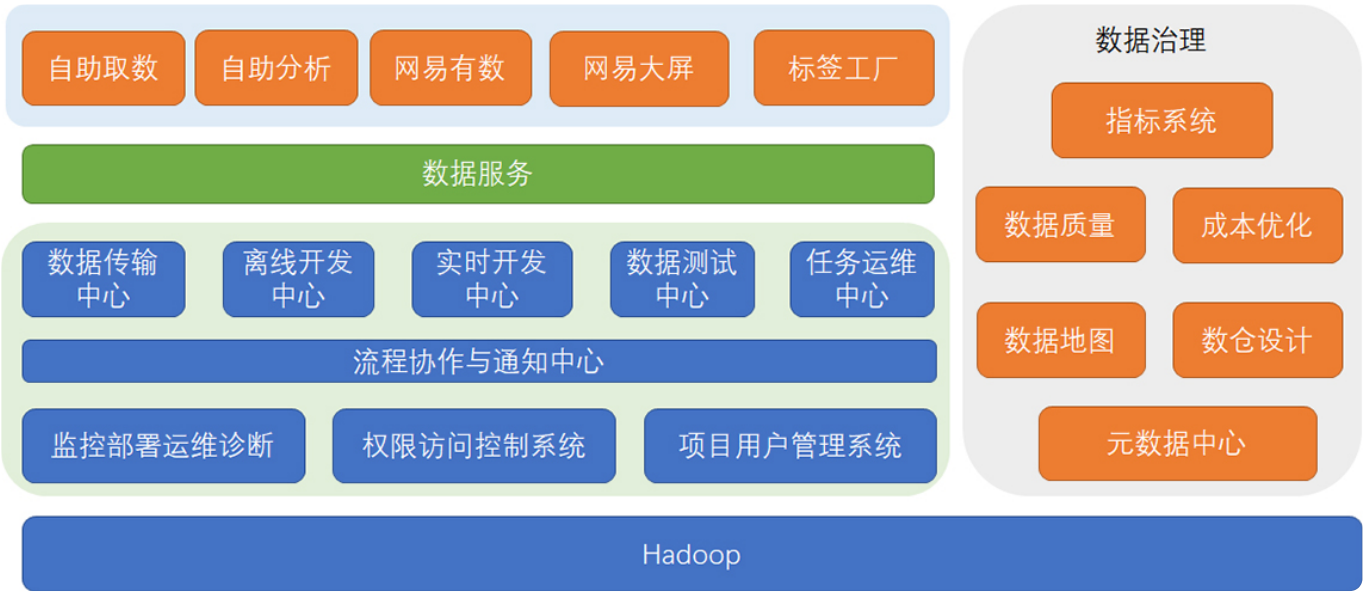
逻辑模型：从用户的视角出发，屏蔽底层的模型设计的实现，面向用户提供逻辑模型。什么是逻辑模型呢？熟悉数据库的同学应该知道，数据库中有一个视图的概念，视图本身并没有真实的数据，一个视图可以关联一张或者多张表，每次在查询的时候，动态地将不同表的查询结果聚合成视图的查询结果。逻辑模型可以类比视图，它可以帮助应用开发者屏蔽底层的数据物理实现，实现相同粒度的数据构造一个逻辑模型，简化了数据接入的复杂度。

性能和稳定性：由于数据服务侵入到用户的访问链路，所以对服务的可用性和性能都有很高的要求，数据服务必须是无状态的，可以做到横向扩展。

OneService 体系的目标是提高数据的共享能力，让数据可以被用得好，用得爽。

数据中台支撑技术

讲完方法论，我们接着要来讲数据中台的支撑技术，因为一个好用的工具，可以让你的数据中台建设事半功倍。



这个图完整地描述了数据中台支撑技术体系，它的底层是以 Hadoop 为代表的大数据计算、存储基础设施，提供了大数据运行所必须的计算、存储资源。

以 HDFS 为代表的分布式文件系统，以 Yarn/Kubernates 为代表的资源调度系统，以 Hive、Spark、Fink 为代表的分布式计算引擎，都属于基础设施范畴。如果把数据中台比作是一个数据工厂，那可以把它们比作是这个工厂的水、电。

在 Hadoop 之上，浅绿色的部分是原有大数据平台范畴内的工具产品，覆盖了从数据集成、数据开发、数据测试到任务运维的整套工具链产品。同时还包括基础的监控运维系统、权限访问控制系统和项目用户的管理系统。由于涉及多人协作，所以还有一个流程协作与通知中心。

灰色的部分，是数据中台的核心组成部分：数据治理模块。它对应的方法论就是 OneData 体系。以元数据中心为基础，在统一了企业所有数据源的元数据基础上，提供了包括数据地图、数仓设计、数据质量、成本优化以及指标管理在内的 5 个产品，分别对应的就是数据发现、模型、质量、成本和指标的治理。

深绿色的部分是数据服务，它是数据中台的门户，对外提供了统一的数据服务，对应的方法论就是 OneService。数据服务向下提供了应用和表的访问关系，使数据血缘可以延申到数据应用，向上支撑了各种数据应用和服务，所有的系统通过统一的 API 接口获取数据。

在数据服务之上，是面向不同场景的数据产品和应用，包括面向非技术人员的自助取数系统；面向数据开发、分析师的自助分析系统；面向敏捷数据分析场景的 BI 产品；活动直播场景下的大屏系统；以及用户画像相关的标签工厂。

这套产品技术支撑体系，覆盖了数据中台建设的整个过程，配合规范化实施，你就可以搭建出一个数据中台，关于具体的细节我会在实现篇中逐一分析讲解，这里你只需要知道这个框架就可以了。

组织架构

我在开篇提到，在网易电商数据中台建设之前，各个部门都会存在一些小的数仓，**那么你有没有想过，为什么会存在这些分散的小数仓？** 归根结底是因为建设这些数仓的人分散在各个业务部门。所以，如果你要建设数据中台，单纯有方法论和支撑技术还不够，还必须要有一个独立于业务部门的中台团队。

数据中台提供的是一个跨业务部门共享的公共数据能力，所以，承担数据中台建设职责的部门一定是一个独立于业务线的部门。这个部门的负责人应该直接向公司的 CTO 汇报工作，当然这个也要取决于数据中台建设的层次，例如在网易内，有云音乐、严选等多个产品线，数据中台的建设层次是在产品级别的，也就是说，云音乐有一个数据中台，严选有一个数据中台，所以严选的数据中台应该向严选的 CTO 汇报。

而独立部门的最大风险是与业务脱节，所以我们对数据中台的组织定位是：**懂业务，能够深入业务，扎根业务。**数据中台要管理所有的指标，而每个业务线之间的指标既有差异，也有交叉，要理解指标的口径定义，就必须要了解业务的过程。同时，当我们要制定一些新的指标时，必须要了解各个业务线新的业务目标，指标的本质还是为业务目标服务的。

综合来讲，什么样的组织架构是适合数据中台建设的呢？

数据产品部门：负责数据中台、数据产品的体系规划、产品设计、规范制定、应用效果跟进，指标口径的定义和维护（有的部门是由分析师管理）。

数据平台部门：负责研发支撑数据中台构建的产品，例如指标系统、元数据中心、数据地图等。

数据开发团队：负责维护数据中台的公共数据层，满足数据产品制定的数据需求。

应用开发团队：负责开发数据应用产品，比如报表系统、电商中的供应链系统、高层看板、经营分析。

而且，中台组织的绩效目标一定是要与业务落地价值绑定的，比如在电商中，我们提供了供应链决策系统，有智能补货的功能，会根据商品的库存，各个地区的历史销售情况，生产加工周期，自动生成补货决策，由人工审核以后，直接推送给采购系统。那我们评估价值时，我们会拿由系统自动生成的采购计划占整体采购计划的比例来衡量数据的应用价值。

最后，数据中台的组织架构改革涉及原有各个部门的利益，所以这个是数据中台构建最难又不得不做的地方，必须要取得高层领导的支持和重视。

课堂小结

这节课，我以自己建设数据中台的经历，带领你了解了数据中台建设的三板斧：方法论、支撑技术和组织架构。在课程的最后，我想再强调几个点。

适合数据中台的组织架构是建设数据中台的第一步，数据中台组织一定是独立的部门，同时要避免与业务脱节，深入业务，要与业务目标绑定。

数据中台支撑技术大规模落地，需要有成熟的系统工具作为支撑，同时要注意这些系统工具之间的联动和打通。

数据中台的方法论可以借鉴，但是不能完全照搬，每个公司的数据应用水平和当前遇到的问题都不相同，可以针对这些问题，分阶段制定数据中台的建设计划，选择性的应用一些技术，例如当前最主要的问题是数据质量问题，那就应该优先落地数据质量中心，提升质量。

最后，让我们回到开篇的那个问题，如何建设数据中台？在我看来，方法论、支撑技术和组织架构，对于建设数据中台，三者缺一不可。而且我想再强调一下，数据中台的建设绝对不是为了建中台而建中台，数据中台的建设一定要结合落地场景，可以先从从一些小的场景开始，但是规划一定是要有顶层设计的。关于具体的操作细节，我会在第二部分，用 8 讲的篇幅来讲给你听。

思考时间

在课程最后，我希望你能够思考一下，哪些数据中台建设的方法论和支撑技术是适合你当前的公司的，如果你们要做数据中台，你所在的组织架构要做哪些变动。

最后，感谢你的阅读，如果这篇文章让你有所收获，也欢迎你将它分享给更多的朋友。

© 版权归极客邦科技所有，未经许可不得传播售卖。 页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇　02 | 关键抉择： 到底什么样的企业应该建数据中台?

下一篇　特别放送 | 史凯：建设数据中台到底有什么用?

精选留言 (15)

写留言

幸福时光
2020-04-06

在传统企业落地数据中台，面临一个很大的挑战是：领导有强烈诉求要做数字化转型，但是具体的执行业务部门基本没有什么数据运营思维，提不出有价值的业务数据问题？像这种情况如何开始数据中台建设？建设好好初步数据中台产品以后，如何运营数据中台业务价值，如何让数据价值落地传统企业？不知老郭后续如何看待这些问题？是否有相关经验分享。谢谢！

展开

作者回复: 我来谈谈我的看法。

依靠业务部门去提数据需求，我觉得你还是想多了，也不太现实，就像你说的，他们根本都没有

数据运营的思维，又怎么能给你提数据问题呢？

这让我想起了我们做数据产品时候的做法， 首先我们的数据产品要深入业务，与业务部门的老大对齐目标。例如在电商业务中，我们跟业务部门的BU老大，对齐，今年的目标主要是两个，一个是要提高爆品产出，一方面降低库存，一方面可以吸粉。另外一个就是要控制毛利。

了解了这个目标以后，我们首先就是要让这个目标可以量化，比如对于爆品的产出，我们需要基于动销率去分析，然后就是持续的跟踪，哪些商品品类动销率比较差，需要提高，哪些商品是0动销，对于0动销的商品，数据产品要进一步分析原因，比如是价格的因素？还是因为商品质量有问题？还是因为商品季节性因素？ 然后给出决策建议，比如下架商品，对商品曝光进行限流，调整商品的价格策略等等，最后，数据产品会自动把这些决策建议推送给业务系统进行执行。

从上面这个案例中，你看到，业务部门不可能给你明确的他们要什么数据，他能给你的是他的业务目标是什么。而数据应用团队，要做的就是对业务目标进行量化，持续跟踪，对于异常要进行诊断分析，给出优化建议，最后一键执行。这个过程最终以数据产品的方式呈现给业务，帮助业务实现数据驱动目标。

在第11节中，我会详细介绍数据应用，相信会对你有帮助。

感谢你的阅读，欢迎你继续跟我交流~

7

Terry郑

2020-04-07

老师您好，我司在探索数据中台的道路上遇到一个问题，就是数据质量不高的问题。
数据质量不高的情况下，很难在数据中挖掘有利价值做数据决策。
在提高数据校验规则以后，各个干系方会想办法做斗争。道高一尺的感觉。
请问如何去从一个角度对数据质量进行切入呢？

展开

作者回复: 我们保证数据问题，做到早发现，早恢复，是依赖数据血缘+稽核监控实现的全链路监控机制，话句话说，如果你加的稽核规则不够多，就可能发现不了问题。那如何来保证稽核规则的完备性的呢？

首先，对于数据中台核心表，数据架构师、主题域的负责人以及对应表的负责人会逐个review 每个表的稽核规则是否完备。

其次，表负责人是表数据质量的第一责任人，我们会对表的数据质量问题进行持续监控，尤其是对下游产生问题的事故进行定责，所以作为第一责任人，他们有动力去完善表的稽核规则。

最后， 稽核规则不可能做到100%的覆盖，只能保证，翻过的错误不要再犯就，所以对于每次事故，我们都会组织复盘，其中重要的一项内容就是补充相关的稽核规则。

通过上述三项措施，可以大幅降低数据质量产生的事故，我可以负责任的说，数据质量不可能说做到100%不出问题，但是可以做到问题不断收敛，犯过的错误不要再犯，这对数据质量来说，已经是极大的改善了。

2

Nick
2020-04-06

看完前几篇的体会之一：必须要拉出两三个数据应用场景，以此为导向再进行数据中台的建设，否则很大可能是无用功。

展开

作者回复: 说的对，没有数据应用场景，数据中台就跟天上的云，始终没有落地的成果，你也讲不清楚，中台到底有啥价值。我觉得，数据应用和数据中台要相互配合构建，可以先从一两个场景开始，滚雪球式的建设。

1

狼迹
2020-04-08

三方面缺一不可，完全赞同郭老师的三驾马车，刚才评论只是想说，项目实施的难点往往是人的因素

作者回复: 对的，人的因素往往最难推动，因为涉及到很多团队的利益，还涉及到能不能找到一些有经验的人来做这个事情。

感谢你的讨论，欢迎你在后面的课程中继续与我在留言区交流~

小熊
2020-04-07

很有收获，期待继续更新

展开

作者回复: 感谢你的支持，希望我的经验对你有所帮助。期待在后续的课程中能够在留言区与你继续交流。

狼迹
2020-04-07

对于数据中台项目的建设，要有远大光荣的理想目标，又要有落地可行的任务计划，最重要是要领导的支持和组织机构来适应，确保目标是一致的，活是大家一起来干的，组织机构放在最后应该是重点也是难点。

作者回复: 我觉得呢，组织架构是我们做这项工作的基础和前提，但是也不是说把这些人放到一起，就能实现数据复用了，还要配合方法论和支撑技术，你要让他能够快速的找到自己想要的表，才能够复用，同时要你按照主题域、分层的方式，让中间加工结果能够落到中台中，被其他的任务复用。而这些，并不是说组织架构在一起就可以了。

所以我给出的结论，是方法论、支撑技术和组织架构三者缺一不可。

lory(侯保国)
2020-04-07

我是一家Saas软件供应商，如果想使用网易的中台服务，并提供给最终的客户，如何使用呢?

作者回复: 感谢你的认可。

我写这个专栏，目的还是希望能够分享网易数据建设的经验，把网易数据中台建设中遇到的坑，如何解决的，分享给大家。

如果你对网易数据中台的相关技术感兴趣，可以搜索网易大数据猛犸。

保护
2020-04-07

看完后收获很大，期待下一篇👍

展开

作者回复: 感谢你的阅读，希望我的经验对你有所帮助，欢迎你继续在留言区与我交流。

Jie
2020-04-06

看专栏有划线记笔记的习惯，老师的文章句句干货，弄的我全文划满线，这不和不划线没区别了么∑(゚Д゚)

作者回复: 希望把最有用的东西呈现给你们，越往后看，就越有料哈，从第4节开始，我们就要进入实践篇啦，这部分的内容会更落地一些，会有比较多的案例，结合一些具体的技术实现，慢慢看~

Samlam
2020-04-06

老师你好。请问用户主题域应该如何设计呢？如何整合不同业务来源的用户数据呢？

作者回复: 你这里面有两个问题。

对于第一个问题，主题域的设计，其实与你所在的企业从事的业务过程密切相关的，你可以理解为它就是业务过程的一个抽象。比如，在物流行业中，会有门店、中转、车队等等，在电商行业中，交易、用户、商品、售后等等，在云音乐中，互动、内容、交易、市场、风控等等。要梳理主题域，你可以先把业务过程中，按照行为事件的方式梳理一下，看看包括哪些业务过程。另外也可以参考一下业务相近的行业对于主题域的划分。还要再说明一下，主题域划分并没有对错，尽可能的覆盖更多表，是主题域划分的一个目标。

如何整合不同业务来源的用户数据。用户在业务中，一般是以维度的方式存在，对于单个业务，需要构建一个统一的用户维表。但是如果涉及多个业务，多个业务用户数据的整合，其实核心技术是id-mapping。

ID-Mapping要解决的核心问题是把相同的用户，在不同应用系统上登录识别成同一个实体，即使是在同一个应用内，同一个人也可能有多重身份，比如未登录和登陆后，可能是两个标识。如何整合这些数据呢，我们把账号和设备的关联关系作为基础，每一个账号在某个设备登录一次，就算一个联系，然后对数据按照权重进行聚类，这个权重就是登录次数，时间等。然后就可以把同

一个设备不同应用，不同设备的相同账号关联起来，识别为同一个实体，这样就实现了不同业务来源数据的整合。

张珮磊想静静
2020-04-06

实际工作中，取得高层领导的支持和重视这方面有什么技巧吗？业务方KPI也很重，数据中台的效果也并不是可以立刻见效的，而且数据中台开发也需要大量的人力，如何说服高层重视并且愿意投入人力，而且可以给出相对充足的时间看最终效果如何

展开

作者回复: 这个问题问到点子上了，问的很好。

一方面是当前业务需求的压力，一方面就是数据中台涉及到大量的模型重构，如何来平衡两者的关系，我的建议是区分团队，专人专办。 单独划出来一拨人，专门做数据中台的数据整合。

至于如何来获得领导的认可。我想，如果你是领导，你最担心是什么？ 最担心就是投了这么多资源，然后效果讲不清楚，钱打了水漂。

我觉得要取得领导的信任，最重要的是要有明确的、可量化的、可持续追踪的目标。然后能够阶段性的验收成果，把风险控制到最低。

数据中台，可以先从一两个典型应用开始，滚雪球式的建设，所以只要有可持续的量化目标和结果，你的领导就可以看到收益，另外，要注重业务成果的展示，而这一部分往往是通过数据应用来实现的，说白了，就是要让这些数据应用为数据中台背书，获得这些数据应用的认可，比如你的需求交付效率确实提高了，你的数据质量确实改善了。

这是我之前使用的一些经验，希望能够对你有所帮助~ 欢迎你继续和我交流。

炬灏
2020-04-06

各种开源技术的支持已经不是挑战，其实面临的最大的挑战是人和业务。

作者回复: 说的对，所以我一直在强调，组织是数据中台的前提，然后才是方法论和支撑技术。

Resilience
2020-04-06

老师，请问下数据中台和数据仓库有啥本质上的区别，感觉数据中台干的这些事情数据仓库也可以办到呀

作者回复: 数据中台，构建于数据湖之上，强调数据只加工一次，数据的复用性，通过数据服务化，提高接口和数据的复用。

传统数据仓库并不是构建于数据湖之上的。而对于现代基于数据湖之上构建的数据仓库，并没有强调数据只加工一次和数据服务化。这就是OneData和OneService方法论。

1

leslie
2020-04-06

个人觉得：设计与开发、维护和优化、整合与平台，这样其实同时作为初期的架构中台架构体系。这块其实目前一直在探索，企业的不同阶段我们应当如何去建立对应的体系设计？
谢谢分享：期待后续的课程。

展开

作者回复: 我觉得，企业不同阶段，数据应用的水平是有差异的，对数据中台的诉求也有不同。

我认为数据应用一般先是从BI 开始的，然后经历数据产品，最后到自助取数。

对于数据中台来说，指标口径一致性问题，数据质量问题，可能是最先面临的问题，然后接下来是效率方面的问题，最后等规模大了，成本问题会更加突出。

所以不同阶段，中台建设的方向优先级是有差异的，以解决当前问题优先。

Randy Liu
2020-04-06

作为之前只关注业务中台建设的，也清楚业务中台与数据中台之间是相互促进，今天看了数据中台的建设思路与内容，有了另一种触动，其实业务中台建设与数据中台建设并非相互促进，而是要预先思考。也就是建设业务中台的时候，还得思考产生的数据能为后续的数据分析，数据统一，数据中台服务建设的契合与准备。

展开

作者回复: 这也是双中台战略的核心。

感谢你的阅读，欢迎你继续和我在留言区互动~